www.sciencemag.org/cgi/content/full/329/5987/75/DC1

Supporting Online Material for

Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude

Xin Yi, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, Xun Xu, Hui Jiang, Nicolas Vinckenbosch, Thorfinn Sand Korneliussen, Hancheng Zheng, Tao Liu, Weiming He, Kui Li, Ruibang Luo, Xifang Nie, Honglong Wu, Meiru Zhao, Hongzhi Cao, Jing Zou, Ying Shan, Shuzheng Li, Qi Yang, Asan, Peixiang Ni, Geng Tian, Junming Xu, Xiao Liu, Tao Jiang, Renhua Wu, Guangyu Zhou, Meifang Tang, Junjie Qin, Tong Wang, Shuijian Feng, Guohong Li, Huasang, Jiangbai Luosang, Wei Wang, Fang Chen, Yading Wang, Xiaoguang Zheng, Zhuo Li, Zhuoma Bianba, Ge Yang, Xinping Wang, Shuhui Tang, Guoyi Gao, Yong Chen, Zhen Luo, Lamu Gusang, Zheng Cao, Qinghui Zhang, Weihan Ouyang, Xiaoli Ren, Huiqing Liang, Huisong Zheng, Yebo Huang, Jingxiang Li, Lars Bolund, Karsten Kristiansen, Yingrui Li, Yong Zhang, Xiuqing Zhang, Ruiqiang Li, Songgang Li, Huanming Yang, Rasmus Nielsen,* Jun Wang,* Jian Wang*

*To whom correspondence should be addressed. E-mail: wangjian@genomics.org.cn (Ji.W.); wangj@genomics.org.cn (Ju.W.); rasmus_nielsen@berkeley.edu (R.N.)

Published 2 July 2010, *Science* **329**, 75 (2010)
DOI: 10.1126/science.1190371

This PDF file includes:

Materials and Methods

Sample description

The 50 ethnic Tibetans analyzed in this study were from two villages in the Tibet Autonomous Region, China. Half of these samples were from the town of Zhaxizhong, Dingri (9 females and 16 males), located at the foot of mountain Jomoglangma (4300 meters in altitude). The remainder were from the town of Zaren, Nachu (13 females and 12 males), which is approximately 250 miles northwest of Lhasa (at 4,600 m). All participants gave a self-report of at least three generations living in the sampling site, and provided informed consent for this study.

The peripheral venous blood samples of 50 ethnic Tibetans was collected using the pipelines dictated by the institutional review board of the Beijing Genomics Institute (BGI). In all subjects, oxygen saturation of blood was measured by Fingertip Oximeter: CMS-50DL twice with thirty minutes of interval. Blood testing was done using standard protocols for the BC-3000 Plus Auto Hematology Analyzer (MINDRAY): erythrocyte quantities were assessed by automated cell counting, and hemoglobin was quantified by spectrophotometry following hemolysis (using the SFT method). Comprehensive medical examinations were also conducted for all individuals during sampling to ensure that only healthy subjects were included in our analysis. All samples and measurements were obtained in the home village of each individual.

DNA extraction, library construction, exome capture and sequencing

Genomic DNA was extracted from the blood samples by the use of QIAamp DNA Blood Mini Kit, according to protocol provided by QIAGEN. Following the manufacturer's protocol, genomic DNA of each individual was hybridized with NimbleGen 2.1M-probe sequence capture array (*S1*) (http://www.nimblegen.com/products/seqcap/) to enrich the exonic DNA in each library. The array is able to capture 18,654 (92%) of the 20,091 genes that has been deposited in Consensus Coding Sequence Region database (http://www.ncbi.nlm.nih.gov/projects/CCDS/).

First, DNA was randomly fragmented by nebulization to an average size of 500bp, and a pair of linkers was ligated to both ends of the resulting fragments. The linker-ligated DNA products were then hybridized to the capture array following NimbleGen's protocol, after which the exome-enriched DNA fragments were eluted from the array and amplified by ligation-mediated PCR, and non-hybridized fragments were then washed out.

Second, the captured DNA fragments were concatenated by DNA ligase and re-sheared to 200bp on average. Thus, we constructed a secondary library from the primary captured DNA library, which enabled the Illumina Genome Analyzer II platform, as previously described (*S2*), with adaptations. We performed sequencing for each captured library independently to ensure each sample had at least ~6-fold coverage. Raw image files were

processed by Illumina Pipeline (version 1.3.4) for base-calling with default parameters and the sequences of each individual were generated as 75bp reads.

Read Mapping and Data quality analysis

Linker or adapter sequences that may be introduced into raw reads during the experiment process were masked before mapping. More concretely, the small portion of adapter and linker sequences within reads was identified by using a local dynamic programming algorithm, and reads that had more than 12 bp overlap with adapter or linker sequences were identified as contaminated reads. The contaminated sequence in reads was then discarded and the remaining sequence was retained. SOAPaligner (*S3, S4*) was used to align the clean reads to the NCBI human genome reference assembly (build 36.3), with a maximum of two mismatches, and parameters set as -a -D -o -r 1 -t -c -f 4. Reads that aligned to the designed target region were collected for SNP identification and subsequent analysis. To evaluate exon capture efficiency, the proportions of reads mapping to target regions and to their flanking regions (within 500 bp) were calculated for each individual. 35.5% of reads mapped to target regions (Table S1) and 68.1% of reads were within 500bp of a target region.

SNP calling and estimation of sample allele frequencies

Calculation of genotype likelihoods

Likelihoods of genotypes of each individual at every genomic location were calculated by SOAPsnp (*S4*). The observed data in site k of a particular read, d_k, contains three elements: (1) o_k: observed allele type; (2) q_k: quality score; (3) c_k: sequencing cycle (coordinate on read) ; and (4) t_k, the t_k-th observation of the same allele from reads with the same mapping location. All three elements in each read are used for the calculation of likelihoods, and sequencing errors are assumed to be independent. The likelihood for genotype S in site k is then

$$p(d_k \mid S) = p(o_k, q_k, c_k \mid S) = p(o_k, c_k \mid S, q_k)p(q_k \mid S)$$

We first estimate a four-dimensional matrix of $p(o_k, c_k \mid S, q_k)$ on a grid of values of o_k, q_k, and c_k for all possible genotypes, based on all of our alignments, using observed mismatch rates. Doing this, we can in effect recalibrate the quality score taking sequencing cycle into account.

$p(q_k|S)$ is the probability of an allele S to have an observation with quality score q_k. The quality distribution of each assumed allele is unknown. Here, we assumed that the distributions from A, C, G, and T are the same; then $p(q_k|S)$ is the function of q_k only.

The same alleles from reads with the same mapping locations were ordered by the sequencing quality scores from low to high. An empirical treatment was used to reduce the quality of the t_k-th observation:

$$q'_k = \theta^{t_k} q_k$$

Here, θ is called a dependency coefficient. The adjusted quality score q'_k, instead of the original q_k, was used in the likelihood matrix. θ is set between 0 and 1. Specifically, $\theta = 0$ means the completely dependent model, and $\theta = 1$ is the completely independent model. A detailed description if this method is provided elsewhere (*S4*).

Allele frequency estimation

Population genetic inferences based on called (inferred) SNPs can lead to serious biases and possibly false inferences if the coverage is not so large that the genotypes are known with absolute certainty for each individual. We have therefore developed a series of statistical techniques that can take uncertainty in genotype calls and allele frequency estimation into account.

To call SNPs and to estimate the allele frequencies in the sample, we use a Bayesian approach which is applied jointly to all individuals. SNP calling based on the joint information from all individuals should be more accurate than SNP calling based on independent analyses of single individuals. The same algorithm which estimates the posterior probability that a SNP is variable can also be used to estimate the frequency on an allele. We will first explain how the algorithm works for a single population. We then subsequently describe how the algorithm works for multiple populations.

Let p_j be the posterior probability that a di-allelic SNP has MAF of $j/2k$, where k is the sample size (number of individuals). We assume that a fraction, p_{var}, of nucleotide sites are variable in the population (not the sample!). Let the observed sequencing data for the SNP be X_i, and let $S = (S_1, S_2, \ldots, S_k)$ be a sample configuration where $S_i \in \Psi$, $\Psi = \{AA, AC, AG, AT, CA, CC, ..., TT\}$. Also, assume that the MAF in the population is p, and let $\chi(S, j)$ be an indicator function which returns 1 if the sample MAF in configuration S is $j/2k$. p_j is then, for $0<i<k$, given by

$$\pi_i = \frac{p_{\text{var}} \sum_{S \in \Psi^k} \left(\chi(S,j) \prod_{i=1}^{k} p(X_i \mid S_i) p(S_i \mid p) \right)}{p_{\text{var}} \sum_{S \in \Psi^k} \left(\prod_{i=1}^{k} p(X_i \mid S_i) p(S_i \mid p) \right) + (1 - p_{\text{var}}) \sum_{S \in \Psi^k} \left(\chi(S,0) \prod_{i=1}^{k} p(X_i \mid S_i) p(S_i \mid p) \right)},$$

and

$$\pi_0 = \frac{\sum_{S \in \Psi^k} \left(\chi(S,0) \prod_{i=1}^{k} p(X_i \mid S_i) p(S_i \mid p) \right)}{p_{\text{var}} \sum_{S \in \Psi^k} \left(\prod_{i=1}^{k} p(X_i \mid S_i) p(S_i \mid p) \right) + (1 - p_{\text{var}}) \sum_{S \in \Psi^k} \left(\chi(S,0) \prod_{i=1}^{k} p(X_i \mid S_i) p(S_i \mid p) \right)},$$

for $i = 0$. $p(X_i \mid S_i)$ is given (up to a scaling factor) by the genotype likelihoods which can be calculated as described above. $p(S_i \mid p)$ can be calculated assuming Hardy-Weinberg equilibrium if the allele frequency p is know. Our algorithm, therefore, proceeds by first estimating p from the raw sequencing reads. The entire calculation can be done very fast using a dynamic programming algorithm for summing over all elements in Ψ^k. In the following we give a detailed description of the algorithmic details of the inference method: We first estimate allele frequencies in each site, and we then estimate the Site Frequency Spectrum (SFS).

Estimating allele frequencies from reads in one site

Let the individuals be $I_1, I_2, \ldots, I_k$, i.e. we assume k individuals.
(1) For each site in each individual, eliminate all reads with Q score < 20. Determine which two nucleotides are most common among {A, C, T, G} and let the set of these nucleotides be B, i.e. if there are 400 A's, 42, C's, 13 T's and 9 G's, then B = {A, C}. Then eliminate all reads that are not elements of B
(2) For i=1 to k

Let n_i be the number of reads of the minor allele in B in individual I_i. Let the total number of reads in B in individual I_i be n_{iT}. Calculate

$$p_i = \frac{n_i - e n_{iT}}{n_{iT}(1-2e)}$$

This is an error corrected estimate of the allele frequency in individual I_i, obtained as the solution for p_i to the equation $n_i = p_i n_{iT}(1-e) + (n_{iT} - p_i n_{iT})e$. The parameter e is the error rate and is considered a fixed parameter, here assumed to be $e = 0.005$. Also calculate $w_i = \frac{2n_{iT}}{n_{iT}+1}$, the inverse of the variance of p_i (up to a scalar).

(3) The estimate of the MAF is then calculated as

$$\hat{p} = \min\{1, p^*\}, \; p^* = \max\left\{0, \sum_{i=1}^{k} p_i w_i \Big/ \sum_{i=1}^{k} w_i\right\},$$

Estimating Sample Allele Frequencies

Likelihood values for all $G \in$ {AA, AC, AG,AT,…,TT} have been calculated using the previously described algorithm. . We are interested in estimating the posterior probability that the minor allele frequency exists in a frequency j in the sample of $2k$ chromosomes. We assume that the prior probabilities of the different genotypes are given by the probabilities predicted under Hardy-Weinberg equilibrium with a MAF of $\hat{p}$. This corresponds to using an empirical Bayesian approach where the shared parameter ($\hat{p}$) first is estimated and then provides a prior for each individual. We will denote the minor allele by 'A' and the major by 'a'. Then a dynamic programming algorithm for calculating the posterior probability is given by (for each site):

If $\hat{p} = 0$, set $p_0 = 1$ and $p_j = 0$ for all $j > 0$.
Else

(1) Set $h_j = 0,\ j = 3,4,\ldots,2k$.

(2) For i=1 to k

Set $P_{AA,} = g_{AA,i}(\hat{p}^2(1-F) + \hat{p}F)$, $P_{Aa} = cf_i g_{Aa,i} 2(1-\hat{p})\hat{p}(1-F)$ and $P_{aa} = g_{aa,i}(1-\hat{p})^2(1-F) + (1-\hat{p})F$.

Here $g_{G,i}$ is the previously calculated likelihood for genotype G in individual i.. The parameter F_i is the inbreeding coefficient and needs to be obtained prior to analyses jointly for all sites We will assume here that $F = 0$.

If i=1

Set $h_0 = P_{aa}$

Set $h_1 = P_{Aa}$

Set $h_2 = P_{AA}$

Otherwise

For $j = 2i$ to 2 (count backwards)

Set $h_j = P_{AA}h_{j-2} + P_{Aa}h_{j-1} + P_{aa}h_j$

Set $h_1 = P_{aa}h_1 + P_{Aa}h_0$

Set $h_0 = P_{aa}h_0$

(3) Set $\pi_j = \dfrac{h_j p_{\text{var}}}{p_{\text{var}}\sum_r h_r + (1-p_{\text{var}})\prod_{i=1}^{k} g_{aa}}, j = 1,2,..,2k$

$$\pi_0 = \frac{p_{\text{var}}h_0 + (1-p_{\text{var}})\prod_{i=1}^{k} g_{aa}}{p_{\text{var}}\sum_r h_r + (1-p_{\text{var}})\prod_{i=1}^{k} g_{aa}}$$

The estimated values of p_i, can then be used for population genetic inferences, either by averaging over p_i, or by using a Maximum *a posteriori* Probability (MAP) estimate of the sample allele frequency. Notice that this procedure explicitly takes into account differences in sequencing depths between samples when estimating allele frequencies, and quantifies the uncertainty in these estimates. Likewise, SNP calling can proceed in a probabilistic fashion by choosing a cut-off for p_0 (p_{2k} is so close to zero that it can be ignored because the definition of p as the minor allele frequency). For example, if we wish to call sites with a probability >95% of being SNPs, we would select all sites with $p_0 < 0.05$.

Extension to multiple populations

We here discuss the extension to two populations, in this case Han (*H*) and Tibetans (*T*). We will use a single estimate of *p*, calculated as previously described for both populations. The main motivation for doing this is to avoid situations in which $\hat{p} = 0$ for one population and $\hat{p} > 0$ in another population. Another justification for using the shared estimate is that we would rather be conservative with regards to inferences of differences in sample allele frequencies between populations. We therefore prefer to use the same prior for both populations.

The joint posterior probability of a site having allele frequency *i* in *H* and *j* in *T*, is then given by

$$\pi_{ij} = \frac{h_j^T h_i^H p_{\text{var}}}{p_{\text{var}} \sum_r \sum_s h_r^T h_s^H + (1 - p_{\text{var}})(\prod_{m=1}^{Tk} g_{aa}^{m,T})(\prod_{m=1}^{Hk} g_{aa}^{m,H})}$$

where T_k is the number of Tibetan individuals and H_k is the number of Han individuals. All functions sub- or super-scripted with either T or H are calculated as previously described marginally for population *T* and *H*, respectively. A SNP is then called if p_{00} is less than some specified cut-off.

Population genetic inferences

Population genetic statistics that do not use linkage/linkage disequilibrium information into account are all functions of Site-Frequency-Spectrum (SFS). In our case, an estimate of the joint SFS for Tibetans and Han is giving by the matrix $p = \{p_{ij}\}$. Statistic such as F_{ST}, the number of segregating sites, the average number of pairwise differences, etc, can be calculated directly from *p* for each gene. This can be done based on the MAP estimate for called SNPs (i.e. SNPs with p_{00} less than some specified cut-off), or it can be done by summing over the values in *p*, thereby taking uncertainty in both SNP calling and inference of allele frequency into account. For example, the number of segregating sites in the Tibetan population in a gene would be calculated as $\sum_{sites}\left(1 - \sum_{i=1}^{2T_k} \pi_{i0}\right)$ and the total number of segregating sites in a gene would be calculated as $\sum_{sites}(1 - \pi_{00})$. Likewise, the average number of pairwise differences per site can be calculated as

$$\frac{\sum_{sites}\left(\sum_{i=1}^{2H_k}\sum_{j=1}^{2T_k} \frac{j(2T_k - j)\pi_{ij}}{\binom{2T_k}{2}}\right)}{\sum_{sites}\left(1 - \sum_{i=1}^{2H_k} \pi_{i0}\right)}.$$

Any other statistic calculated on a per site basis, which normally for a single variable site with sample allele frequency i in H and j in T is given by $f(i, j)$, can similarly be calculated as

$$\frac{\sum_{sites}\left(\sum_{i=1}^{2H_k}\sum_{j=1}^{2T_k} f(i,j)\pi_{ij}\right)}{\sum_{sites}\left(1-\sum_{i=1}^{2H_k}\pi_{i0}\right)}.$$

Our inference of natural selection is primarily based on a new statistic aimed at detecting strong changes in allele in one population. Pairwise differences in allele frequencies can be quantified using F_{ST}. We use the F_{ST} estimator of Reynolds *et al.* (*S5*), based on the MAP estimates for SNP frequencies where sites are considered if they satisfy $(1.0\text{-}p_{00}) > 0.01$. We also excluded sites for which the minor allele has MAP frequency estimate of 0 in at least two populations or for which no data was available for the Danish population. . We then use the classical transformation by Cavalli-Sforza (*S6*),

$$T = -\log(1 - F_{ST})$$

to obtain estimates of the population divergence time T in units scaled by the population size. For each RefSeq gene, we calculate this value between the Tibetans and Han populations (T^{TH}), and between these populations and a Danish population (T^{TD} and T^{HD}), for which data obtained using similar techniques was previously published for 200 individuals, providing very accurate estimates of allele frequencies. The length of the branch leading to the Tibetan population since the divergence from Han, is then obtained as

$$PBS = \frac{T^{TH} + T^{TD} - T^{HD}}{2}$$

A population's *PBS* value represents the amount of allele frequency change at a given locus in the history of this population (since its divergence from the other two populations). This approach is similar to the "locus-specific branch lengths" statistic used by Shriver *et al.* (*S7*), except that by using the above log-transformation, we obtain additive distances that place branches of different magnitudes on the same scale. This statistic should be very powerful to detect selection. It should have power, for example to detect incomplete selective sweeps, a type of selection that is highly relevant here and which most other statistics based on the SFS have little power to detect.

Evaluation of the PBS statistic

Recent simulation studies have shown that F_{ST}-based statistics (*S8*) have more power to detect recent adaptation when selection is acting on standing variation. Because of the very short divergence time between Han and Tibetan individuals, and the fact that the waiting time to a new mutation might be large, we expect much local adaptation to be

driven by selection acting on standing variation rather than *de novo* mutations. The test statistic we are using is, therefore, a simple transformation of F_{ST} designed to take advantage of an outgroup and to identify Tibetan specific selection.

To evaluate if this approach also has power to detect selection on *de novo* mutations, we performed a small scale simulation study. Using the Wright-Fisher model simulator *sfscode* (*S9*), we simulated 3 populations (representing Danes, Tibetans and Han) introducing one new selected mutation in the Tibetan population at the time of the split of Han and Tibetans. We simulated data sets under a range of scaled selection coefficient, ($g = 2Ns$, where s is the selection coefficient and N is the population size), assuming the population size $N = 1000$ for each of the 3 populations, and we assumed divergence times between Danes and Asians, and Han and Tibet, of 1680 and 120 generations, respectively. The locus size was set to 1kb, the population mutation rate and the population recombination rate were set to 0.001 per site. No further complications to the demographic model were used in this analysis, because these simulations were only used for evaluating the power of the *PBS* statistic, and not to generate *P* values for empirical observations.

Often the selected mutation in a simulation will be lost after a few generations due to the effect of genetic drift. However, as we were interested in evaluating the power under a complete or incomplete selective sweep, we only examined simulation replicates where the selected mutation was not lost from the Tibetan population, effectively conditioning on the presence of the allele. To determine critical values, we ran neutral simulations (no selected mutation was introduced). The power was then calculated by comparing the simulations with selection to simulations without selection. For comparison, we also calculated Tajima's *D* (*S10*) for each simulation replicate and evaluated the power of Tajima's *D* based on the same simulations. Results indicated that *PBS* has strong power to detect a recent selective sweep (Figure S3). The power of Tajima's *D*, in contrast, is quite low in this setting, potentially due to low numbers of segregating sites. Because our exonic data contains relatively few SNPs per gene, the high power of *PBS* under these conditions represents an important advantage for our analysis. A similar set of simulations were also conducted with the population recombination rate elevated to 0.01, in order to simulate a locus that is ten times longer, but with only the same number of sampled sites (analogous to our exonic data). Results were qualitatively similar: *PBS* retained high power in these simulations, while Tajima's *D* had modestly higher power than it had with shorter loci (data not shown).

Demographic estimation and neutral simulations

For the inference of demographic parameters we used the unfolded site frequency spectrum (based on ancestral alleles shared by chimpanzee and macaque genomes) of the synonymous sites (61,347 SNPs) in the Han and Tibetan samples. Parameter inference was carried out with the software package ∂a∂i (version 1.2.3) (*S11*). We took ancestral population events such as the out-of-Africa bottleneck from the model inferred by Gutenkunst *et al.* (*S11*), but we estimated parameters pertaining to the two Asian samples studied here. Models were compared via Akaike and Bayesian Information Criteria; the

best fitting-model is shown in Figure S2. As further detailed in the legend of this figure, this model involves a population split 2,750 years ago. The Han size is initially small but grows larger, while the Tibetan size is initially large but contracts with time. Migration occurs from the Tibetan sample to the Han sample, but 20% of the Tibetan gene pool is replaced by Han admixture at the present time. A wide variety of models were tested, but the model shown in Figure S2 fit better, for example, than the same model with symmetric migration, and much better than a similar model lacking the ancestral African time and growth estimates of Gutenkunst *et al.* (*S11*). The model of European history from Gutenkunst *et al.* (*S11*) was used for the history of the Danish sample in the simulations described below.

Neutral simulations under the model estimated above were used to calculate *P* values for the *PBS* values inferred for each gene in the ethnic Tibetan sample. Simulations were run using the program *ms* (*S12*) with demographic parameters from the above model and recombination rates drawn randomly from the map of McVean *et al.* (*S13*). Gene lengths for simulations were sampled randomly from the lengths of all human genes. One million simulations were performed for each number of SNPs (for 1 to 15 SNPs) or using 5-SNP bins (from 20 to 40 SNPs) and conservatively comparing genes to simulations with slightly fewer SNPs (*e.g.* comparing a gene with 28 SNPs to simulations with 25 SNPs). *P* values were defined simply as the proportion of simulated replicates yielding a higher *PBS* value than empirically observed for a particular gene.

Genotyping and association testing for a candidate SNP at EPAS1

The SNP at *EPAS1* showing the most dramatic frequency difference between ethnic Tibetan and Han samples (located at position 46441523 on chromosome 2) was genotyped in a larger sample of 366 ethnic Tibetans (from the same localities, and collected via the same protocols, as described above. Genotyping was done by use of the mass-spectrometry-based MassArray platform of Sequenom (San Diego, CA, USA). PCR and extension primers were designed using Assay Design v3.1 (Sequenom, San Diego, CA, USA). Forward and reverse PCR primers were ACGTTGGATGTCCATGTCTGACCCTTCCAC and ACGTTGGATGTATTGTGAGGAGGGCAGTTG. Genotyping primers had the unextended sequence GACCCTTCCACGCCTGT, extending to a “C” or “G” for the alternate alleles.

PCR reactions were performed in 5µl PCR cocktail mix, consisting of 1µl DNA template (10-25 ng/µl), 1 × PCR Buffer (including 2 mmol/L MgCl2), 2 mmol/L MgCl2, 500 µmol/L dNTP mix, 0.1pmol/µl of each PCR primer, and 0.5U Hotstar Taq (Roche). PCR conditions were as follows: incubation at 94°C for 15 min, followed by 45 cycles of 20 sec at 94°C, 30 sec at 56°C, 1 min at 72°C, and a final extension of 3 min at 72°C. Shrimp alkaline phosphatase treatment was performed to dephosphorylate unincorporated dNTPs under the following conditions: 37°C for 40 min, 85°C for 5 min, cooling to 4°C.

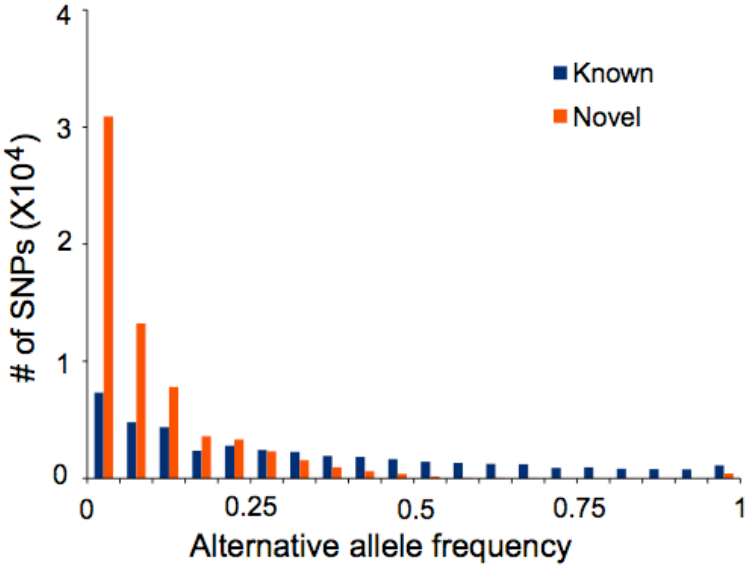
The iPLEX primer extension reaction was performed using the iPLEX cocktail mix (Sequenom, San Diego, CA, USA), which contains buffer, iPLEX termination mix,

iPLEX enzyme and extension primers, under the following conditions: the DNA sample is denatured at 94°C, Strands are annealed at 52°C for 5 seconds and extended at 80°C for 5 seconds ,The annealing and extension cycle is repeated four more times for a total of five cycles and then looped back to a 94°C denaturing step for 5 seconds and then enters the 5 cycle annealing and extension loop again. The five annealing and extension steps with the single denaturing step are repeated an additional 39 times for a total of 40. A final extension is done at 72°C for three minutes and then the sample is cooled to 4°C. Six milligram clean resin was added into 384-well PCR plate to desalt the iPLEX extension products before mass spectrometric analysis. An average of 3-10 nl products were dispensed onto a 384-element SpectroCHIP bioarray (Sequenom) by a nanodispenser. MassARRAY Workstation version 3.4 software (Sequenom) was used to process and analyze iPLEX SpectroCHIP bioarrays. Positive and negative control samples were run at each step and on each chip.

Association testing was performed using simple linear regressions of the measurements oxygen saturation, erythrocyte count, and hemoglobin concentration on the genotypes of the focal SNP at *EPAS1*. The genotypes were encoded as numerical values 0, 1, 2 corresponding to homozygous, heterozygous and homozygous (for the other allele) genotypes. We used the model $E[Y|X_i] = \beta_0 + \beta_1 X_i$ and tested whether the slope (β_1) is different from zero. Here Y is the quantitative trait and X_i takes the values {0, 1, 2} for the genotypes at SNP site *i*. The regressions were performed for the full sample of 366 individuals, for each of the two villages separately (Table S5). The analysis made use of the linear regression function from the R programming language and F-test *P*-values were recorded. Genotypes at the focal *EPAS1* SNP were uncorrelated with gender. To further control for gender-related phenotypic differences, we also performed association testing in females only, and in males only. Results were very similar to the overall results: associations for erythrocyte count and hemoglobin quantity remained statistically significant or marginally significant, and associations for oxygen saturation did not approach statistical significance. Since population stratification may be an issue, we calculated the inflation factor from non-associated SNPs (*S14*) in the full sample, and used this inflation factor to compute *EPAS1* association *P*-values corrected for population stratification for our most differentiated SNP. The results remained statistically significant. The phenotypic associations observed for focal SNP at *EPAS1* were also compared against 48 additional SNPs from around the genome, genotyped in the same large sample. The *P* value observed for *EPAS1* was a clear outlier from this set (Fig. S5). Positions for these "genomic control" SNPs were as follows: chr1-12491677, chr1-27151140, chr1-45846284, chr1-52675081, chr1-53448299, chr1-65630810, chr1-110567989, chr1-154829060, chr1-194962365, chr1-201404410, chr2-43955048, chr2-71215412, chr2-178202419, chr2-218391384, chr3-19936334, chr3-57113459, chr4-77284346, chr5-35912031, chr5-96357847, chr5-172274635, chr5-172274640, chr6-25881584, chr6-26164595, chr6-133146813, chr6-151715282, chr7-6032976, chr8-101796892, chr8-105430170, chr10-29931167, chr11-1934128, chr11-3642019, chr11-61767439, chr11-74793531, chr11-89541802, chr11-106702850, chr12-9208040, chr14-64267910, chr17-39581008, chr17-64702135, chr19-1005255, chr19-1776926, chr19-6864707, chr19-8097184, chr19-46314107, chr19-59665806, chr20-33678648, chr20-36217869, chr22-40816669.

Figure S1. Alternative site frequency spectra (SFS) for Tibetan exome data.
a) Comparison between AFS of known (blue; in dbSNP v129) and novel SNPs (red).
b) Comparison between AFS of empirical data and the estimated demography model.

a)



b)

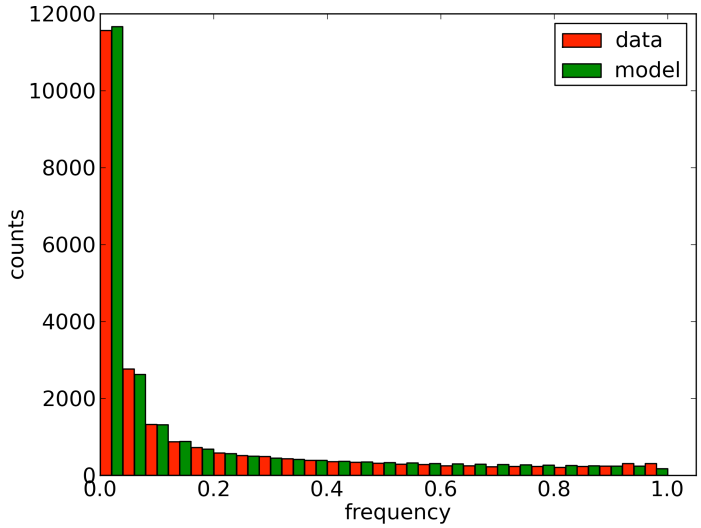


Figure S2. Illustration of best-fitting demographic model according to Akaike Information Criterion and Bayesian Information Criterion. Parameters in red were estimated; parameters in black were fixed according to the model of Gutenkunst *et al.* (2009). Estimates for inferred parameters were as follows: The ancestral non-African population grows to a size of N_{AS} = 7360 at time T_1 = 42,955 years ago (all time estimates assume 25 years per generation). At time $T2$ = 2,750 years ago, the Han and Tibetan lineages split, with the Han population having initial size N_H = 288 and the Tibetan population having initial size N_T = 22,642. At time T_3 = 1,973 years ago, the Tibetan population begins exponential decline to a final size of N_{TF} = 1,270. At time T_4 = 1,813 years ago, the Han population begins exponential growth to a final size of N_{HF} = 12,850, and migration from the Tibetan to the Han population occurs at rate m_{HT} = 0.00526. Finally, at the present time, a proportion F_{TH} = 0.2 of the Tibetan gene pool is drawn from the Han sample (instantaneous admixture).

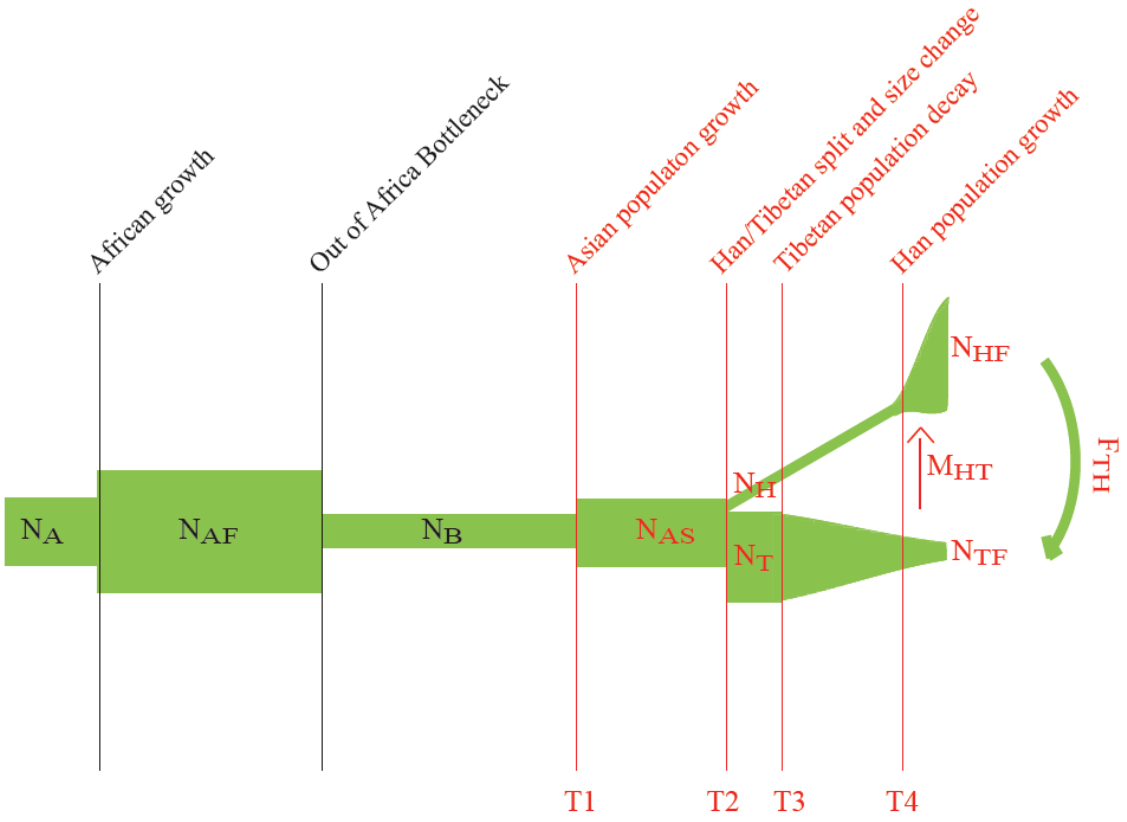


Figure S3. Power of the *PBS* and Tajima's *D* statistics to detect a recent selective sweep, depending on the strength of selection (X-axis). Simulations were conducted as described in the Supplemental Text.

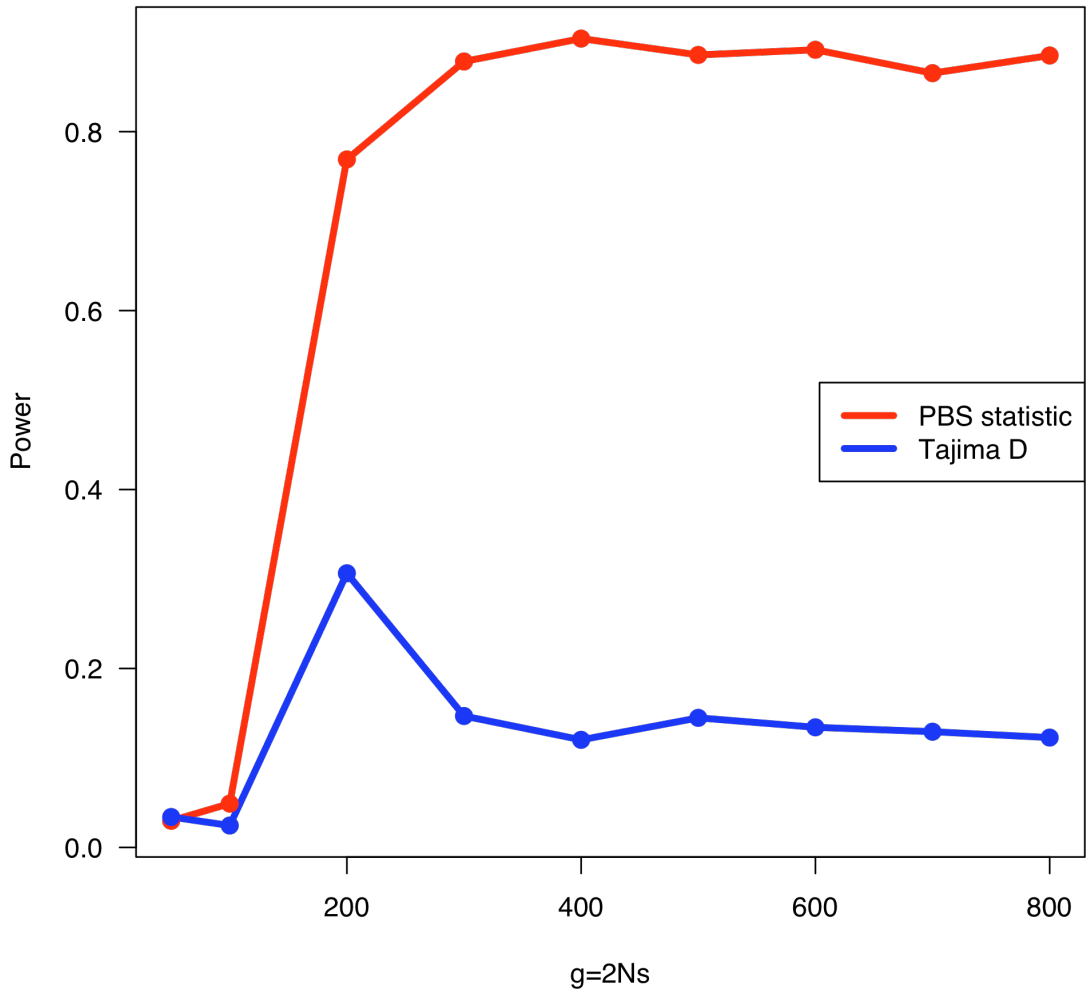


Figure S4. Distribution of $\log_e$ p-values for 48 genomic control SNPs regressed against erythrocyte count, and for the genotyped *EPAS1* SNP (red star)

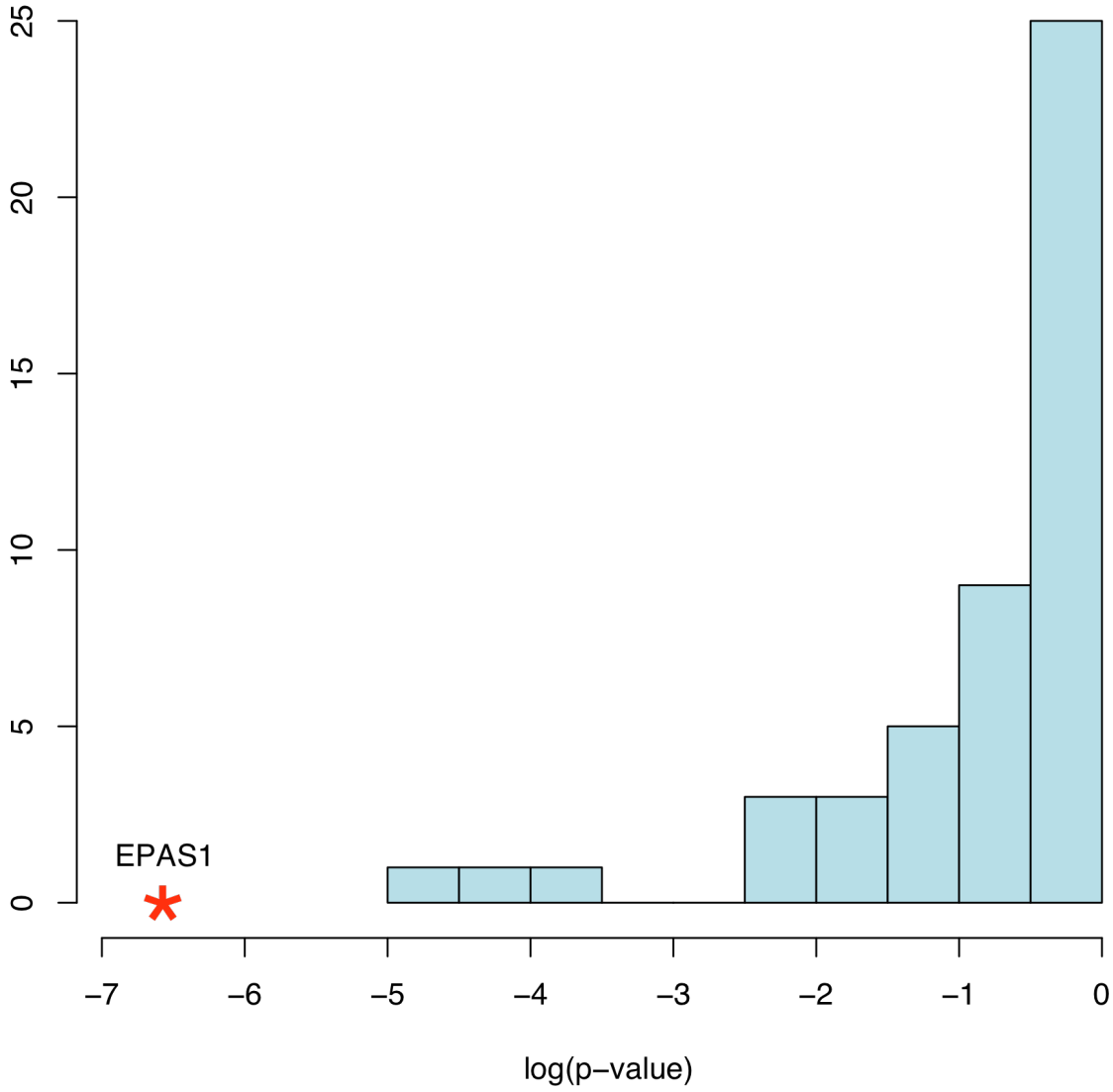


Figure S5. Linked pairs or groups of genes that appear on the list of most extreme *PBS* values (Table 2) are shown in green. Nearby candidate genes are marked in red.

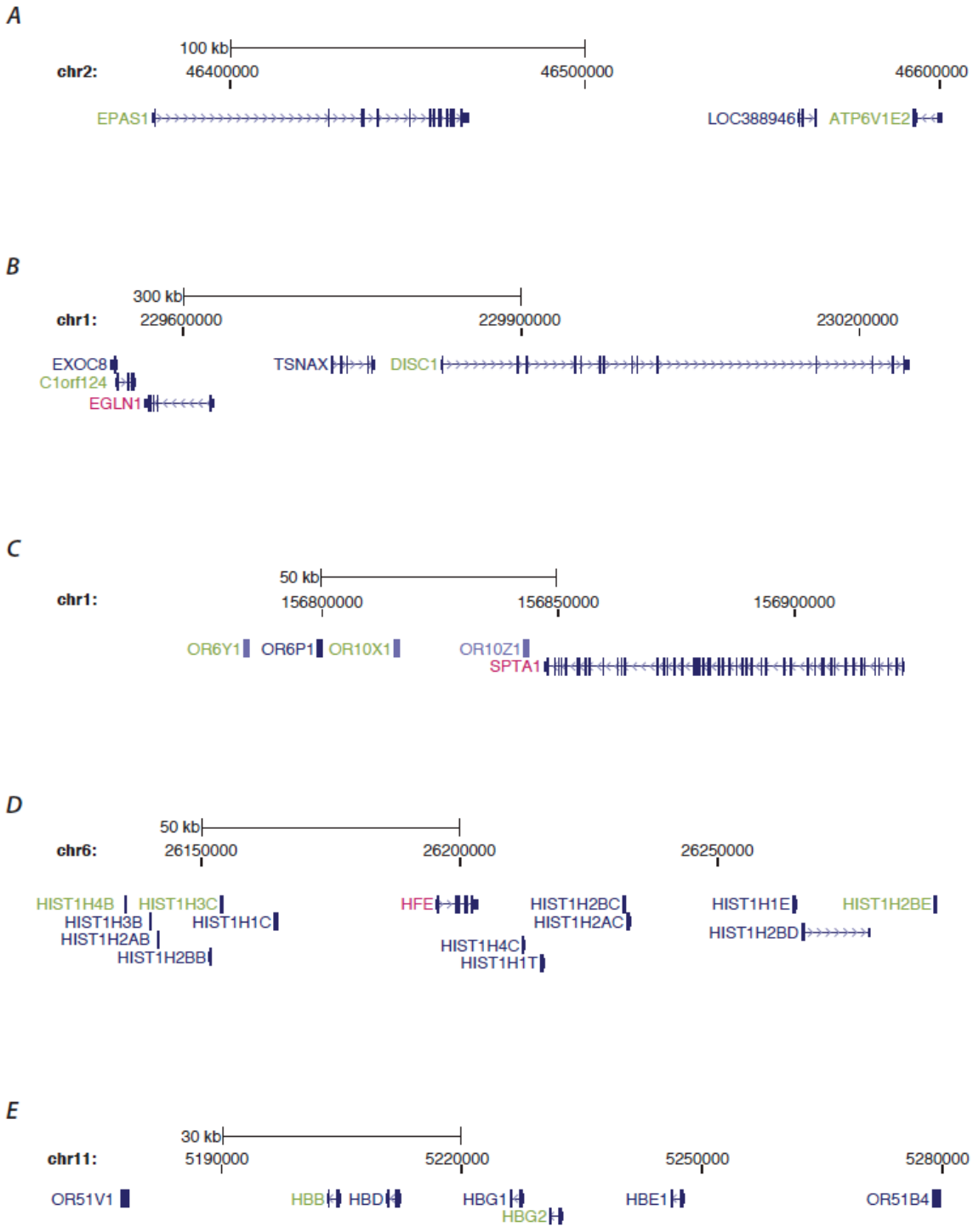


Table S1. Data production by individual sample

Sample ID	Raw reads (x1e+6)	Raw data yield (Mb)	Reads mapped to genome (x1e+6)	Reads mapped to target region (x1e+6)	Data mapped to target region (Mb)	Mean depth of target region	Coverage of target region (%)	Average read length (bp)	Rate of nucleotide mismatch (%)
DR-F11	44.75	3,356	13.7	4.08	201.92	5.92	87.01	62.07	2.04
DR-F14	35.51	2,663	31.38	11.15	614.64	18.02	95.25	69.93	0.96
DR-F17	38.23	2,867	32.23	11.6	638.52	18.72	95.26	70.32	0.9
DR-F18	47.41	3,556	18.88	6.48	316.87	9.29	90.13	61.47	2.17
DR-F19	42.30	3,172	12.37	4.02	199.88	5.86	88.05	62.32	2.03
DR-F35	44.19	3,314	29.94	11.92	596.22	17.48	97.55	61.43	1.29
DR-F40	33.01	2,476	29.07	10.4	557.68	16.35	96.37	69.78	1.08
DR-F6	31.84	2,388	28.48	10.08	556.31	16.31	94.94	69.98	0.96
DR-F8	46.18	3,463	19.08	6.7	328.47	9.63	92.75	61.55	2.24
DR-M10	39.01	2,926	34.29	12.85	704.01	20.64	95.97	69.76	1
DR-M19	40.06	2,404	34.51	8.06	398.73	11.69	96.21	61.5	0.76
DR-M22	41.24	1,526	21.79	9.54	440.34	12.91	97.09	56.16	1.26
DR-M23	21.06	1,685	17.91	7.34	424.31	12.44	95.1	73.92	0.89
DR-M24	40.76	3,057	36.47	14.18	785.87	23.04	96.05	70.68	0.85
DR-M28	40.58	3,043	27.67	11.63	578.14	16.95	97.89	60.69	1.21
DR-M30	41.19	3,089	28.34	11.82	587.35	17.22	97.95	60.64	1.21
DR-M31	36.41	2,731	31.49	10.95	587.69	17.23	95.51	69.66	1.1
DR-M38	40.10	3,007	25.97	12.1	642.27	18.83	96.9	65.67	1.1
DR-M42	40.14	3,010	26.38	10.77	536.19	15.72	97.39	60.96	1.28
DR-M43	40.80	3,060	25.85	10.52	526.3	15.43	97.4	61.27	1.1
DR-M44	40.91	1,514	22.67	9.41	435.23	12.76	97.22	56.28	1.1
DR-M46	37.95	2,846	16.9	7.47	372.13	10.91	96.55	60.77	1.18
DR-M7	45.35	3,401	17.43	5.88	298.79	8.76	90.37	64.51	1.84
DR-M8	46.58	2,795	39.98	11.85	597.59	17.52	97.14	62.83	0.72
DR-M9	41.28	3,096	27.96	11.49	573.37	16.81	97.63	60.92	1.22
NQ-F15	66.40	4,980	47.15	15.59	819.98	24.04	96.01	66.33	1.15
NQ-F16	38.01	2,851	30.67	10.97	603.04	17.68	96.71	69.86	0.99

NQ-F17	65.31	4,898	44.84	13.82	717.99	21.05	96.92	65.33	1.21
NQ-F19	70.48	3,947	48.48	15.33	819.98	24.04	95.38	67.93	1.17
NQ-F20	41.22	1,525	22.94	8.88	411.69	12.07	97.14	56.51	1.04
NQ-F24	62.88	4,716	49.48	18.6	977.9	28.67	97.68	65.93	1.52
NQ-F25	64.36	4,827	45.8	16.34	845.22	24.78	97.15	65.19	1.26
NQ-F26	59.85	4,489	48.25	18.59	978.24	28.68	97.53	65.85	1.56
NQ-F32	67.11	5,033	47.13	14.32	753.46	22.09	94.68	66.49	1.13
NQ-F34	46.40	3,480	15.25	4.97	242.51	7.11	90.47	61.21	2.11
NQ-F35	57.04	4,278	43.41	15.01	794.74	23.3	97.61	66.08	1.14
NQ-F36	58.83	4,412	39.95	13.27	689.68	20.22	97.27	64.98	1.29
NQ-F7	58.93	4,420	40.16	12.39	647.73	18.99	97.44	65.41	1.23
NQ-M12	48.98	3,674	19.23	5.8	294.7	8.64	89.62	64.17	1.76
NQ-M13	39.68	2,976	33.82	13.64	758.92	22.25	92.71	70.03	0.99
NQ-M20	64.05	4,803	47.25	16.55	875.91	25.68	96.39	66.44	1.3
NQ-M21	46.41	3,481	18.97	5.47	268.1	7.86	87.6	61.04	2.15
NQ-M26	41.56	3,117	28.48	10.24	524.59	15.38	96.27	64.15	1.58
NQ-M31	39.38	1,457	17.59	7.51	362.24	10.62	96.66	59.19	0.98
NQ-M32	62.55	4,691	49.82	19.97	1043.05	30.58	97.75	65.45	1.59
NQ-M33	58.68	4,401	47.42	15.04	788.25	23.11	97.08	65.29	1.41
NQ-M35	55.80	4,185	47.4	17.67	978.92	28.7	97.06	70.58	0.87
NQ-M5	59.33	4,450	46.85	17.76	925.37	27.13	97.87	64.84	1.61
NQ-M7	59.36	4,452	47.69	15.65	816.22	23.93	96.62	65.26	1.44
NQ-M9	63.90	3,578	43.68	10.55	551.2	16.16	96.73	65.39	1.25

Table S2. Variation detection from Tibetan exomes

SNP discovery for functional classes of sites

Genomic features		Known	Novel	Total
		# of SNPs	# of SNPs	# of SNPs
CDS	synonymous	14,439	12,312	26,751
	nonsynonymous	11,421	26,634	38,055
	nonsense	73	541	614
Intron		14,547	23,623	38,170
5'UTR		848	1,129	1,977
3'UTR		895	1,100	1,995
Intergenic		15	16	31

Table S3. Additional statistics for the 30 genes with highest Tibetan *PBS* values.

Gene	refseq ID	S Tibetan	∏ Tibetan	S Han	∏ Han	T_{TH}	T_{TD}	T_{HD}
EPAS1	NM_001430	12.64	0.10	8.86	0.17	0.57	0.70	0.24
C1orf124	NM_032018	3.37	0.16	4.15	0.27	0.14	0.53	0.12
DISC1	NM_018662	18.31	0.13	11.99	0.16	0.16	0.49	0.15
ATP6V1E2	NM_080653	1.02	0.24	3.00	0.19	0.12	0.50	0.12
SPP1	NM_001040060	6.15	0.18	4.17	0.28	0.13	0.59	0.25
PKLR	NM_000298	8.91	0.10	5.04	0.20	0.06	0.85	0.45
C4orf7	NM_152997	3.43	0.24	2.11	0.10	0.20	0.27	0.01
PSME2	NM_002818	4.15	0.12	3.79	0.17	0.09	0.56	0.21
OR10X1	NM_001004477	5.11	0.21	5.04	0.37	0.10	0.49	0.15
FAM9C	NM_174901	4.00	0.07	2.22	0.21	0.14	0.36	0.07
LRRC3B	NM_052953	3.81	0.14	1.08	0.23	0.19	0.25	0.00
KRTAP21-2	NM_181617	3.00	0.45	3.11	0.22	0.26	0.23	0.07
HIST1H2BE	NM_003523	2.39	0.09	2.41	0.20	0.09	0.62	0.29
TTLL3	NM_001025930	7.84	0.08	6.38	0.19	0.09	0.56	0.24
HIST1H4B	NM_003544	3.71	0.15	5.02	0.13	0.12	0.43	0.14
ACVR1B	NM_004302	4.09	0.19	3.81	0.19	0.13	0.29	0.03
FXYD6	NM_022003	2.07	0.20	2.01	0.07	0.18	0.24	0.04
NAGLU	NM_000263	5.11	0.13	3.60	0.10	0.16	0.23	0.02
MDH1B	NM_001039845	6.23	0.19	6.36	0.20	0.07	0.61	0.31
OR6Y1	NM_001005189	3.10	0.32	2.08	0.46	0.10	0.34	0.08
HBB	NM_000518	2.32	0.39	2.14	0.47	0.08	0.46	0.17
OTX1	NM_014562	3.57	0.18	2.38	0.18	0.12	0.30	0.05
MBNL1	NM_207292	6.96	0.17	3.75	0.08	0.18	0.18	0.01
IFI27L1	NM_206949	3.14	0.25	2.55	0.11	0.18	0.18	0.01
C18orf55	NM_014177	7.95	0.17	4.68	0.11	0.15	0.24	0.03
RFX3	NM_134428	6.24	0.16	4.47	0.07	0.20	0.16	0.00
HBG2	NM_000184	2.47	0.17	1.46	0.06	0.17	0.17	0.00
FANCA	NM_000135	40.40	0.08	33.13	0.23	0.11	0.62	0.39
HIST1H3C	NM_003531	2.47	0.23	2.05	0.35	0.05	0.72	0.43
TMEM206	NM_018252	2.22	0.16	0.68	0.04	0.17	0.16	0.00

Table S4: Population frequencies and mean phenotypes at the focal *EPAS1* SNP

Allele/genotype	Tibetan frequency	Han frequency	Danish frequency	mean hemoglobin concentration	mean erythrocyte count	mean oxygen saturation
C	0.13	0.9125	1	n/a	n/a	n/a
G	0.87	0.0875	0	n/a	n/a	n/a
CC	10	n/a	n/a	178	5.3	87.5
CG	84	n/a	n/a	178.9	5.6	86.68
GG	272	n/a	n/a	167.5	5.2	86.42

Table S5. Association testing *P* values for the focal *EPAS1* SNP, for the full sample and for each village separately (Dingri and Naqu). For phenotypes with significant *P* values, regression coefficients (β_1), standard errors, and sample sizes (n) for the linear regressions are also given.

Sample	SaO2 *P*	Erythrocyte *P*	Erythrocyte β_1	Erythrocyte SE	Erythrocyte n	Hemoglobin *P*	Hemoglobin β_1	Hemoglobin SE	Hemoglobin n
All Tibetans	0.726	0.00145	-0.236	0.0734	314	0.00127	-9.23	2.84	358
Dingri only	0.805	0.00188	-0.284	0.0901	198	0.00458	-9.14	3.19	240
Naqu only	0.467	0.0609	-0.214	0.113	116	0.00166	-13.6	4.21	118

Table S6. Population genetic statistics for selected *a priori* candidate genes for altitude adaptation in the Tibetan sample.

Gene	refseq ID	Description	S Tibetan	∏ Tibetan	S Han	∏ Han	T_{TH}	T_{TD}	T_{HD}	PBS Tibetan
NOS3	NM_000603	nitric oxide synthase 3 (endothelial cell)	26.07	3.37	17.76	3.22	0.01	0.09	0.06	0.02
HIF1A	NM_181054	hypoxia-inducible factor 1, alpha subunit	13.65	0.73	6.13	0.25	0.02	0.01	0.02	0.01
MB	NM_203377	myoglobin	6.44	2.94	6.27	2.61	0.02	0.04	0.09	-0.02
ACE	NM_000789	angiotensin I converting enzyme isoform 1	23.00	3.94	18.15	3.06	0.02	0.08	0.18	-0.04
CYP11B2	NM_000498	cytochrome P450, subfamily XIB polypeptide 2	13.43	2.83	12.44	2.47	0.03	0.14	0.31	-0.07

Supplemental References and Notes

S1. T. J. Albert *et al.*, Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903-905 (2007).

S2. D. R. Bentley *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**, 53-59 (2008).

S3. R. Li, Y. Li, K. Kristiansen, J. Wang, SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714 (2008).

S4. R. Li *et al.*, SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**: 1124-1132 (2009).

S5. J. Reynolds, B. S. Weir, C. C. Cockerham, Estimation of coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**: 767-779 (1983).

S6. L. L. Cavalli-Sforza, Human Diversity. *Proc. 12th Int. Congr. Genet.* **2**, 405-416 (1969).

S7. M.D. Shriver *et al.*, The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics* **1**:274-286 (2004).

S8. Y. Kim, H. Innan, Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics* **179**, 1713-1720 (2008).

S9. R. D. Hernandez, A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*. **24**, 2786-2787 (2008).

S10. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

S11. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**, e1000695 (2008).

S12. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).

S13. G. A. McVean *et al.*, The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581-584 (2004).

S14. S.-A. Bacanu, B. Devlin, K. Roeder, Association Studies for Quantitative Traits in Structured Populations. G*enetic Epidemiology* **22**, 78–93 (2002)